BIOINFORMATICS APPLICATIONS NOTE

Vol. 23 no. 4 2007, pages 517–519
doi:10.1093/bioinformatics/btl615

Genetics and population analysis

Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome

Winston Lau[†], Tai-Yue Kuo[†], William Tapper, Simon Cox[1] and Andrew Collins[*]

Human Genetics Division, Duthie Building (Mailpoint 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK and [1]Southampton Regional e-Science Centre, School of Engineering Sciences, University of Southampton, Southampton SO17 1BJ, UK

Received on September 21, 2006; revised on November 14, 2006; accepted on November 28, 2006
Advance Access publication December 1, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Summary: Linkage disequilibrium (LD) maps increase power and precision in association mapping, define optimal marker spacing and identify recombination hot-spots and regions influenced by natural selection. Phase II of HapMap provides ~2.8-fold more single nucleotide polymorphisms (SNPs) than phase I for constructing higher resolution maps. *LDMAP-cluster*, is a parallel program for rapid map construction in a Linux environment used here to construct genome-wide LD maps with >8.2 million SNPs from the phase II data.

Availability: The LD maps, *LDMAP-cluster* and documentation are available from: http://www.som.soton.ac.uk/research/geneticsdiv/epidemiology/LDMAP

Contact: arc@soton.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Linkage disequilibrium (LD) describes the tendency of alleles at markers in close proximity to be inherited together more frequently than expected under random segregation. Precise characterization of LD structure underpins efficient mapping of disease genes by association. Maniatis *et al*. (2002) developed an analogue to linkage maps in centimorgans with maps expressed in LD units (LDUs), which have ~1500-fold higher resolution (Tapper *et al*., 2005), and lengths reflecting the number of generations since an 'effective' bottleneck (Zhang *et al*., 2004). Improved localization and substantial increases in power are found when disease mapping with LDU maps (Maniatis *et al*., 2005).

The *LDMAP* program constructs LD maps from single nucleotide polymorphism (SNP) data in population samples using the 'interval' algorithm (Maniatis *et al*., 2002). The program constructs LD maps from either phase unknown (genotypic) data or phase-known (haplotypic) data. Further details of the core methodology are given in Supplementary material. Map construction is computationally intensive employing composite likelihood to estimate a parameter, epsilon (ε), describing the decline of association in each interval between adjacent SNPs.

Phase II of HapMap (International HapMap Consortium, 2005), provides ~2.8-fold more SNPs than phase I. The huge volume of data imposes a considerable computational burden addressed here through the implementation of a parallel algorithm, in the program *LDMAP-cluster*, deployed on a Linux Beowulf cluster. We have used this program to construct genome-wide LDU maps from phase II data for the four HapMap populations. A detailed description of the data are given in the Supplementary materials.

2 IMPLEMENTATION

LDMAP-cluster is written in C, as a wrapper program that encapsulates *LDMAP*. We deployed the program on a Linux Beowulf cluster of over 900 processors. The batch queuing and job management is administrated by Open-PBS (Portable Batch System), http://www.openpbs.org/.

The segment-based parallel approach is illustrated in Figure 1. We established that assembly of maps in segments of ~2000 SNPs loses minimal information and provides substantial reductions in computing time (Supplementary Figure 1). We also examined the effect on map quality of varying the number of pairwise observations used to estimate epsilon in each map interval. An optimum 'interval window' of informative SNP pairs separated by no more than ~100 intervals was identified (Supplementary Figure 2). Map segments are submitted and constructed as individual jobs on the cluster. The parallel processing is accomplished by the concurrent submission of all segments.

LDMAP-cluster is a 64 bit program, enabling access to more memory than conventional 32 bit platforms. The program features synchronous processing supporting multiple SNP dataset submissions. To efficiently utilize dual-processor machines in the cluster, segments are assigned as two jobs per submission. In addition to job monitoring commands (i.e. '*showq*' and '*qstat*') supplied by Open-PBS, a custom-made program, '*checkSeg*', tracks the status of the submitted jobs grouped by SNP dataset.

A segment of 2000 SNPs requires 5–10 h of computation (AMD Opteron 2 GHz with 2 GB RAM), corresponding to the minimum time for construction of the whole map given complete parallelization.

LDMAP-cluster is compatible with a Linux Beowulf cluster with Open-PBS installed as the batch scheduler. Recompilation of the program is essential for linking to the platform specific libraries. Minor modification of the code responsible for job submission is required for porting onto a Linux cluster with a different batch scheduler. Compatibility across all platforms is difficult to

[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

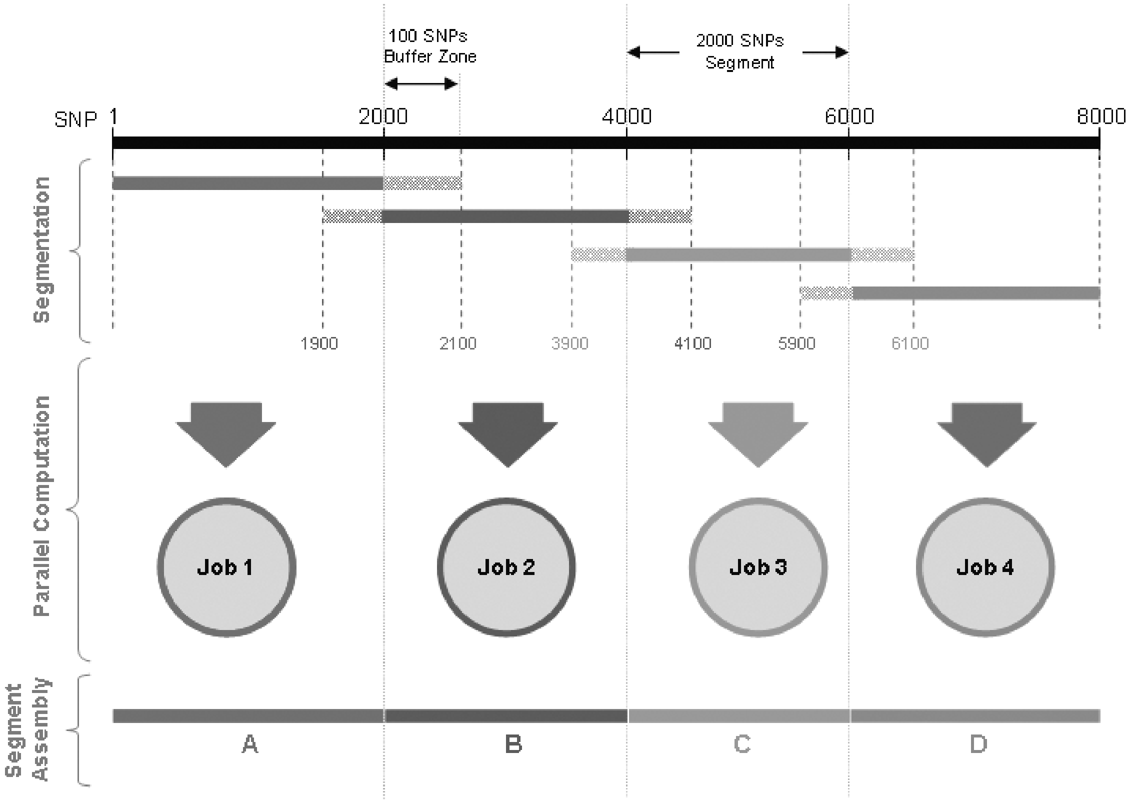

Fig. 1. A chromosome is divided into segments of ~2000 SNPs. A 'buffer zone' of 100 SNPs extends from the ends of each segment to minimize loss of information. Buffer zones are eliminated in map assembly and segments are connected end to end to form the complete map.

guarantee given different hardware (e.g. 32 or 64 bit), software (e.g. PBS or Condor) and administrative environments (e.g. versions of glibc and Tcl/Tk libraries), but modification for local systems should be straightforward as the software is written in standard C. Further technical issues are discussed in detail in the Supplementary materials and supporting website.

3 RESULTS

Tapper *et al*. (2005) describe a genome-wide LD map constructed from ~490 k SNPs (post-screening) from HapMap phase I public release #16 for the CEU population. We describe here maps from all four HapMap populations with 1.9–2.3 million SNPs per population. These data were analyzed in 4195 segments of ~2000 SNPs. Approximately 8.2 million SNPs were processed in ~25 170 computing hours achieved over about one month real-time. The phase II LD maps resolve ~31% of the 'holes' (intervals constrained to the upper limit of three LDUs, Service *et al*., 2006) in the phase I maps where the LD structure is not fully characterized. Such regions are more frequent in large outbred populations, such as those represented in HapMap, where recombination events have accumulated in narrow regions over many generations creating locally high-haplotype diversity. Considering the hugely increased marker density the relatively small proportion of resolved holes suggests that many holes correspond to particularly intense recombination hot-spots. Disease gene mapping by association is expected to be particularly difficult in these areas (Service *et al*., 2006).

Although the broad pattern of LD is consistent between the two HapMap phases (Fig. 2), the fine scale structure of steps and blocks differs in many regions. Increasing SNP density recovers structural details from regions with lower marker coverage in phase I but differences also reflect changes in the sequence build and the resolution of some holes, (which may locally increase or decrease map length).

Overall the phase II maps are 3.1% longer (Table 1), a modest increase consistent with the essentially additive property of the LDU map distances noted previously (Ke *et al*., 2004).

4 DISCUSSION

Genome-wide LDU maps constructed using *LDMAP-cluster* have substantially higher marker density than maps published for the CEU population (Tapper *et al*., 2005). The maps should guide marker selection, empower genome-wide association studies and facilitate other genomic studies. The LD pattern at fine scale is described by these maps, and applications to disease association mapping are expected to increase power and precision for localization of disease genes, consistent with existing evidence (Maniatis *et al*., 2005). The LD pattern is highly consistent between the high-resolution (HapMap release #20) and low-resolution (release #16) maps, despite small differences in overall map length attributable to changes in the sequence and the better characterized LD structure.

Efforts are now underway to generate large case-control and other phenotype samples for association studies with many thousands of

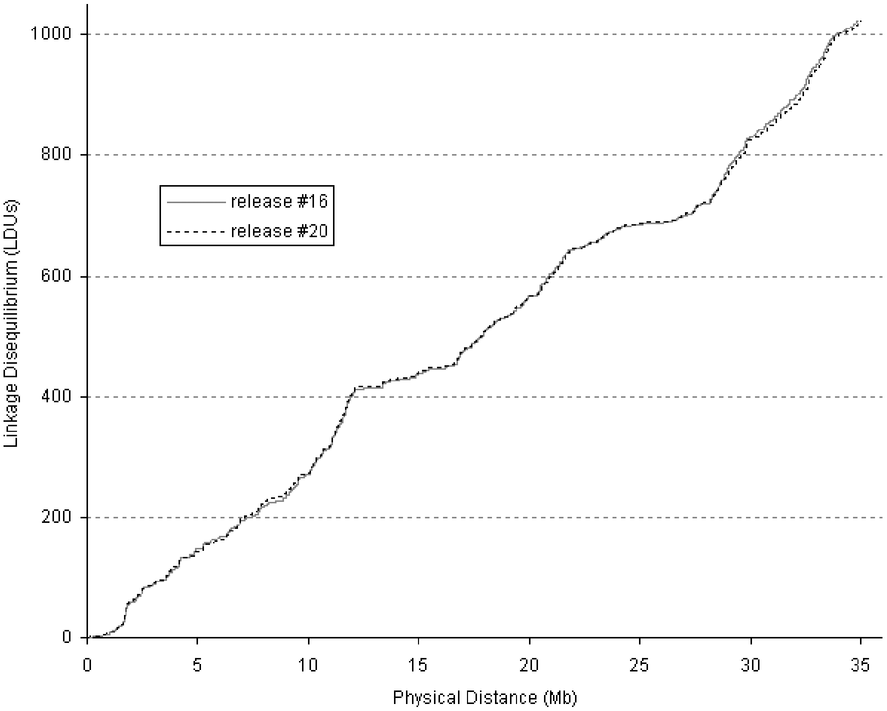

Fig. 2. LD maps of chromosome 22 (CEU) constructed from HapMap #16 (13 959 SNPs) and #20 (26 721 SNPs). The LD pattern is highly consistent between the two HapMap phases.

Table 1. Characteristics of the LDU maps

Populations	CEU	CHB	JPT	YRI	Σ
No. of holes in LD map					
Phase I release #16	2911	4879	3731	2979	14 500
Phase II release #20	2033	3838	2900	1216	9987
Diff.	−30%	−21%	−22%	−59%	(avg.) −31%
Overall LD map length (in LDUs)					
Phase I release #16	56 250	62 686	56 655	79 499	255 091
Phase II release #20	57 819	64 930	58 730	81 345	262 826
Diff.	+2.8%	+3.6%	+3.7%	+2.3%	(avg.) +3.1%

SNPs. The complexities of processing and analyzing such huge bodies of data are an area of rapid research. We anticipate that the genome-wide LDU maps and software tools developed will facilitate association mapping in these samples and contribute to studies of recombination, selection and population history. Applications to data from other organisms, including a recent application to the Bovine genome (Khatkar *et al*., 2006), demonstrate the wide-applicability and utility of this form of genetic map for describing and analyzing LD structure with high-resolution.

ACKNOWLEDGEMENTS

This research is supported by a University of Southampton e-Science centre Postgraduate Research grant.

Conflict of Interest: none declared.

REFERENCES

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Ke,X. *et al*. (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet*., **13**, 577–588.

Khatkar,M.S. *et al*. (2006) A first generation metric linkage disequilibrium map of bovine chromosome 6. *Genetics*, **174**, 79–85.

Maniatis,N. *et al*. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA*, **99**, 2228–2233.

Maniatis,N. *et al*. (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. *Hum. Mol. Genet*., **14**, 145–153.

Service,S. *et al*. (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat. Genet*., **38**, 556–560.

Tapper,W. *et al*. (2005) A map of the human genome in linkage disequilibrium units. *Proc. Natl Acad. Sci. USA*, **102**, 11835–11839.

Zhang,W. *et al*. (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl Acad. Sci. USA*, **101**, 18075–18080.